清华大学数据挖掘课程幕课习题

第一章

第二节

1.“教育不是灌输，而是点燃火焰” 这一思想出自于：苏格拉底。
2.如何学好数据挖掘技术？认真学习幕课视频；充分利用课后阅读材料；勤于动手，实践出真知；主动思考，知其然，知其所以然。

第三节

1. 在超市环境中对客户位置轨迹进行记录和分析的主要目的有哪些？对拥挤人群进行预警；优化商场布局；个性化营销。
2. 在实际数据分析工作中，数据类型转换和数据自身的错误是面临的主要挑战之一。正确
3. 大数据和传统数据分析相比，核心特征就是数据量大。错误，是数据量，数据类型和数据产生的速度。

第四节

1.理想的数据挖掘工作成果应当：Interesting；useful；hidden。
2.ETL 系统主要包括：数据提取；数据转换；数据装载。

第五节

1. 分类器在训练样本上的学习误差越低越好。这是错误的，如果误差越小，说明在空间中分类的线越复杂，对于新样本的判断不一定准确。就像一个死读书的人，在面对他没有见过的知识时，不知道是一个道理，因此，也不是越大就越好。
2. 混淆矩阵中 False Negative 的含义是：被错误的分为负类的样本。
3. 在 ROC 分析中，分类器的性能曲线的理想状态是：越靠上越好，AUC 趋近于 1.
4. 以下最有可能涉及代价敏感分类问题的是：银行信用卡评分模型。
5. 假设目标客户占人群的 5%，现根据用户模型进行打分排序，取 1000 名潜在客户中排名前 10%的客户，发现其中包含 25 名目标客户，问此模型在 10%处的提升度是多少？
6.
解析；假如 100 个客户，只有八个人对商品感兴趣，传统方法就是打 100 个电话，才会知道到底是那八个人。用数据分析方法，对用户进行建模，把用户接受产品的可能性算出来，把前百分之十的用户拿出来，其中可能占了真实感兴趣的百分之四十的用户，百分之四十除

以百分之十，就是四，也就是所谓的提升度。
这道题：目标客户占人群百分之五，则目标客户=1000*0.05=50
前百分之十中真实目标客户占比为：25/50=0.5
提升度= 0.5/0.1=5

第六节

1. 聚类与分类的主要区别在于：
(1 分)
数据维度不同
数据类型不同
数据有无标签
计算复杂度不同
2. 线性回归模型由于自身的局限性只能描述变量间的线性关系。
(1 分)
正确
错误

第七节

1. GPU 作为高性能计算设备的优点包括：
(1 分)
低成本
高计算密度
可独立使用
安装便捷
多选 2. 有效的数据挖掘工作需要哪些因素的支持？
(1 分)
高质量的数据
合适的算法模型
强悍的计算平台
丰富的领域知识
单选 3. 在互联网时代，个人隐私信息的现状是：
(1 分)
没有人知道我是一条狗
只要自己平时注意，还是妥妥的
可能偶尔会泄露
裸奔，必须的
单选 4. 如何才能最有效采集到用户可能不愿公开提供的信息？
(1 分)
晓之以理，动之以情
威逼利诱，瞒天过海

隐匿用户身份
随机问卷题目
判断 5. 在隐私保护的问卷调查中，针对两个互补问题，用户也可用 Yes/No 回答，与用 True/False 本质上是一样的。
(1 分)
正确
错误
单选 6. 以下哪条描述体现了并行计算的思想?
(1 分)
子又生孙，孙又生子，子又有子，子又有孙，子子孙孙无穷匮也
人多力量大，众人划桨开大船
书山有路勤为径，学海无涯苦作舟
不积跬步，无以至千里，不积小流，无以成江海
单选 7. 云计算的核心特征是什么?
(1 分)
Pay As You Go
看云卷云舒
服务器集群
网络化服务
单选 8. 云计算领域中的 SaaS 指的是:
(1 分)
平台即服务
基础设施即服务
软件即服务

第八节

1. 短期股票价格波动难以精准预测的主要原因在于现有模型本身不够精密。
(1 分)
正确
错误
单选 2. 彩票号码难以预测的原因在于:
(1 分)
数据样本不够大
号码的纯随机性
现有模型不够复杂
特征维度太高
单选 3. 以下哪条描述最贴近幸存者偏差现象?
(1 分)
成王败寇
盲人摸象
赢者通吃
真理往往掌握在少数人手里
单选 4. 两个变量 X 和 Y 呈现负相关性，说明:

（1 分）
X 增大会导致 Y 减小
X 减小会导致 Y 增大
X 增大不会导致 Y 增大
然而并不能说明什么
单选 5. 知名企业的 CEO 中身材高大者的比例高于人口平均水平，这是因为：
（1 分）
身材高大者智商高
身材高大者情商高
身材高大者工作更努力
身材高大者容易树立威信
单选 6. 针对数据挖掘领域，以下哪个观点是正确的：
（1 分）
数据=财富
You cannot be too careful!
算法为王
神秘莫测，高不可攀

第二章

第一节

多选 1. 以下关于数据预处理的描述正确的是：
（1 分）
需要借助领域知识
核心内容就是缺失数据填充
数据挖掘工作的基础性工作
主要靠标准化算法自动处理
单选 2. 小张的个人信息中身份证号倒数第二位是单数，性别为女。这种情况被称为：
（1 分）
Missing Data
Inconsistent Data
Noisy Data
Redundant Data
单选 3. 学生小明在调查问卷中没有回答下述问题：“你去年的工资收入和前年相比是否有所增加？”对这种情况最恰当的描述是：
（1 分）
完全随机缺失
N/A
数据未提供
异常数据
单选 4. 以下针对缺失值问题的阐述正确的是：

（1 分）
删就一个字
用均值填充即可
用中位数填充即可
具体问题具体分析

单选 5. 某大一男生体检数据中体重值缺失，相对合理的填充值是：
（1 分）
40 公斤
60 公斤
80 公斤
100 公斤

单选 6. 假设男生用 1 表示，女生用 0 表示，某人的性别未填，应该如何处理？
（1 分）
填 1
填 0
填均值 0.5，必须的
可根据其它信息（如身高、体重）推测

单选 7. 以下关于离群点（Outlier）和异常点（Anomaly）关系的论述正确的是：
（1 分）
一回事，说法不同而已
离群点一定是异常点
异常点一定是离群点
不能简单判定

第二节

1. 关于离群点的判定：
（1 分）
主要看其与近邻的平均距离
主要看其与近邻的最大距离
需要考虑相对距离因素
主要靠感觉

单选 2. 采用 LOF 方法进行离群点检测时：
（1 分）
LOF 值越小越疑似离群点
LOF 值越大越疑似离群点
LOF 值越接近 1 越疑似离群点
LOF 值越接近 0.5 越疑似离群点

单选 3. Case A：两人名字不同，身份证号相同。 Case B：两人同名同姓，身份证号不同。
（1 分）
A 为重复数据可能性大
B 为重复数据可能性大
我读书少，看不出什么区别

单选 4. 在记录手机号码的时候，相对而言：

(1 分)
前三位不容易记错
末尾三位不容易记错
中间三位不容易记错
都一样

单选 5. 在记录英语国家人名时：
(1 分)
姓容易写错
名容易写错
没有明显区别

单选 6. 对英语国家的人群而言：
(1 分)
姓的区分度大
名的区分度大
没有明显区别

第三节

单选 1. 按 A, B, C, D 打分的考试成绩数据属于：
(1 分)
数值型（连续）
数值型（离散）
序数型（Ordinal）
标称型（Nominal）
字符串

多选 2. 在对标称型数据（如颜色、职业等）进行编码时：
(1 分)
按 1,2,3,4...顺序编码即可
类别较少时，可考虑采用扩维法
不同编码可能会影响数据的空间分布
不好处理，删了算了

多选 3. 在大数据分析中，利用采样技术可以：
(1 分)
降低获取数据的成本
减少需要处理的数据量
有助于处理不平衡数据
提高数据的稳定性

单选 4. 对于极度不平衡的二分类数据集，应特别注意：
(1 分)
整体的准确率
多数类样本的准确率
少数类样本的准确率

两类样本准确率的均值

单选 5. SMOTE 的工作原理是：

（1 分）

对多数类样本进行下采样

对少数类样本进行克隆复制

对少数类样本通过插值进行上采样

对整体样本进行随机采样

第四节

单选 1. 很多人感觉到自己的收入与官方公布的平均收入相去甚远，最有可能的解释是：

（1 分）

自己工作不够努力，怨不得别人

统计样本不具有代表性

个体收入分布极度不均衡

错觉，都是错觉

多选 2. Pearson’s product moment correlation coefficient 可用来判断：

（1 分）

X 和 Y 是否正相关

X 和 Y 是否负相关

X 和 Y 是否不相关

X 和 Y 之间的因果关系

单选 3. 在 Box Plots 当中，一个盒子越扁说明在该维度上：

（1 分）

25%到 75%之间的数据分布较为集中

25%到 75%之间的数据分布较为分散

离群点较少

离群点较多

单选 4. 适合可视化高维数据的方法是：

（1 分）

圆饼图

散点图

平行坐标

直方图

单选 5. 数据可视化工作：

（1 分）

锦上添花，可有可无

不学就懂，一看就会

主要用于展示最终结果

贯穿数据挖掘工作全过程

第五节

1. 熵衡量的是系统的不确定性，熵值越大（接近于 1）说明系统的不确定性越低。
（1 分）
正确
错误
单选 2. 假设某数据集的原始熵值为 0.7， 已知某属性的信息增益为 0.2，那么利用该属性进行划分后数据集的熵值为：
（1 分）
0.9
0.7
0.5
0.2
单选 3. 以下方法中可以确保获得最优属性子集的是：
（1 分）
Top K Individual Features
Sequential Forward Selection
Sequential Backward Selection
Simulated Annealing
Exhaustive Search
单选 4. 关于分支定界法不正确的描述是：
（1 分）
树状搜索算法
随机搜索算法
依赖属性的单调性假设
能够减少搜索空间
多选 5. 进行属性选择的原因是：
（1 分）
属性可能存在冗余
属性可能存在噪声
降低问题复杂度
个人喜好

第六节

1. 特征选择与特征提取的关系是：
（1 分）
特征提取包含特征选择
特征选择包含特征提取
一码事，说法不同而已
It is like comparing apples and oranges.
单选 2. 平面图中的老鹰能够被人们识别的原因是：
（1 分）

体积大
为人所熟知
长得有个性
观察角度合适
单选 3. 在 PCA 变换中，应尽量把数据向什么方向投影：
(1 分)
数据集中的方向
数据散布大的方向
数据分组特征明显的方向
平行于原始坐标轴的方向
单选 4. PCA 变换中不包含以下哪一种操作：
(1 分)
去均值
矩阵特征值分解
属性值标准化
坐标变换
单选 5. 假设样本数大于维数，利用 PCA 技术，可以把 N 维数据降到：
(1 分)
只能到 1 维
只能到 N-1 维
1 到 N-1 维
取决于样本的类别数

第七节

1. 如果将 PCA 应用于带标签的分类数据：
(1 分)
程序直接崩溃
效果杠杠的
驴唇不对马嘴
视情况而定
单选 2. LDA 与 PCA 最本质的区别是：
(1 分)
能够降到的维数不同
计算效率不同
降维的目标不同
我读书少，看不出来
单选 3. 当样本个数小于数据维数的时候，LDA 不能正常工作的原因是：
(1 分)
类间散布矩阵不满秩
类内散布矩阵不满秩
计算量过高
Fisher 准则无意义
单选 4. 当类中心重合的时候，LDA 不能正常工作的原因是：

（1 分）
Fisher 准则函数分母为零
类内散布矩阵奇异
Fisher 准则函数恒等于零
类间散布矩阵满秩
单选 5. 对于二分类问题，LDA 只能将原始数据降到 1 维的原因是：
（1 分）
类间散布矩阵秩为 1
类内散布矩阵秩为 1
原始数据维度过高
原始数据维度过低
单选 6. 关于 LDA 和 PCA 投影方向描述正确的是：
（1 分）
必然相同
必然不同
LDA 总是优于 PCA
世事难料

第三章

第一节

1. 有监督的学习和无监督的学习的根本区别在于：
（1 分）
学习过程是否需要人工干预
学习样本是否需要人工标记
学习结果是否需要人工解释
学习参数是否需要人工设置
单选 2. 已知池中有两种鱼，比例为 7:3，若随机捞上一条，按照 70%和 30%概率随机猜测其种类，则整体误差最接近于：
（1 分）
20%
30%
40%
50%
单选 3. 2015 年 10 月，中国共产党第十八届中央委员会第五次全体会议公报指出：坚持计划生育基本国策，积极开展应对人口老龄化行动，实施全面二孩政策。提问：小明的妈妈有两个孩子，已知其中一个是男孩儿，问另一个也是男孩儿的概率是：
（1 分）
二分之一
三分之一

四分之一
真的不关我的事
单选 4. 已知甲乙丙三人射击命中率分别为 0.8，0.6 和 0.5，若每人各开一枪，则目标被命中的概率最接近：
（1 分）
0.85
0.90
0.95
1.00
单选 5. 当化验报告呈阳性的时候，正确的做法是：
（1 分）
心如死灰，万念俱灭
散尽家财，及时行乐
置若罔闻，我行我素
及时复检，防止假阳性

第二节

1. 朴素贝叶斯分类器的朴素之处在于：
（1 分）
只能处理低维属性
只能处理离散型属性
分类效果一般
属性之间的条件独立性假设
单选 2. 以下关于两个变量 X 和 Y 说法正确的是：
（1 分）
若独立一定不相关
若不相关一定独立
若独立不一定不相关
我已经晕了
单选 3. 两个事件 A 和 B 条件独立指的是：
（1 分）
P(A, B)=P(A)P(B)
P(A, B)=P(A|B)P(B)
P(A|B, C)=P(A|C)
P(A|B)=P(A)
单选 4. 以下关于拉普拉斯平滑说法正确的是：
（1 分）
防止计算条件概率时分母为零
防止计算条件概率时分子为零
用于解决训练集中的噪声
用于解决训练集中的异常值
单选 5. 在文本分类应用中，关于词袋模型的描述正确的是：
（1 分）

任何一个单词只能存在于某一个词袋中
一个单词可能存在于多个词袋中但频率不同
所有词袋中单词的并集就等同于词汇表
词袋模型描述的是单词在所有文本中出现的频率

第三节

1. 作为一种分类器，决策树模型的主要优点是:
(1 分)
训练时间短
可解释性好
善于处理缺失值
鲁棒性好
单选 2. 下列哪一种情况被称为过学习现象:
(1 分)
在训练集上 A 优于 B，在测试集上 A 也优于 B
在训练集上 A 优于 B，在测试集上 B 优于 A
相对于分类数据集，决策树过于简单
在训练集上决策树的误差很小
单选 3. 任何一个候选属性在生成的决策树中:
(1 分)
必须被使用
只能被使用一次
可以被使用多次
可以在任意位置被使用多次
单选 4. 以下关于决策树的说法正确的是:
(1 分)
决策树越复杂，分类能力越强
在性能相同的情况下，通常选择能充分利用各种属性的决策树
对于某一个数据集，只有一个决策树可以将其完美分开
对于某一个数据集，可以生成多个决策树
多选 5. 奥卡姆的剃刀指的是:
(1 分)
Entities are not to be multiplied beyond necessity.
Among competing hypotheses, the one with the fewest assumptions should be selected.
The simplest explanation is usually the correct one.
中世纪英国上流社会的一种生活用品。

第四节

1. 为什么一般不推荐在决策树中使用“生日”属性：
(1 分)
星座信息更有说服力
容易造成过学习
可能的取值太多，计算量过大
两个人可能生日相同

单选 2. 决策树模型中建树的基本原则是：
(1 分)
取值多的属性应放在上层
取值少的属性应放在上层
信息增益大的属性应放在上层
应利用尽可能多的属性

多选 3. 哪些情况下必须停止树的增长：
(1 分)
当前数据子集的标签一致
没有更多可用属性
当前数据子集为空
当前训练误差已经较低

单选 4. 关于决策树剪枝操作正确的描述是：
(1 分)
从中间节点开始
从叶节点开始
有助于保持树的平衡
可以有效降低训练误差

单选 5. 在决策树模型中，校验集的用途是：
(1 分)
用于校验模型的训练误差
用于校验模型的测试误差
用于校验模型的正确性
用于控制对模型的剪枝操作

单选 6. 决策树模型中应如何妥善处理连续型属性：
(1 分)
直接忽略
利用固定阈值进行离散化
根据信息增益选择阈值进行离散化
随机选择数据标签发生变化的位置进行离散化

第四章

第一节

1. 如图所示的感知机（阈值为 0）实现的逻辑功能是：
（1 分）
或门
与门
非门
与非门

单选 2. 在感知机的判决函数中，w0 的作用是：
（1 分）
为了后续学习算法推导的方便
其实在实际中可以略去
控制判决平面到原点的距离
控制判决平面的方向

单选 3. 我们很难刻意忘掉一个人的原因是：
（1 分）
记性好，没办法
刻骨铭心，矢志不渝
天长地久有时尽，此情绵绵无绝期
神经元的大规模分布式信息存储机制

第二节

1. 若神经元的误差对某输入的权重的偏导大于零说明：
（1 分）
权重应增加
权重应减小
不能确定

单选 2. 根据 Delta 规则，在 stochastic learning 模式下，若神经元的实际输出大于期望输出，权重应：
（1 分）
顺势而为：增大
反其道而行之：减小
若相应输入大于零：减小
若相应输入小于零：减小

单选 3. 以下关于感知机说法正确的是：
（1 分）
在 batch learning 模式下，权重调整出现在学习每个样本之后
只要参数设置得当，感知机理论上可以解决各种分类问题

感知机的训练过程可以看成是在误差空间进行梯度下降
感知机的激励函数必须采用门限函数
单选 4. 下图所示真值表对应的逻辑电路是:
(1 分)
或非门
与非门
异或门

第三节

1. 采用 Sigmod 函数作为激励函数的主要原因是:
(1 分)
有固定的输出上下界
计算复杂度较低
导数存在解析解
处处可导
单选 2. 以下关于感知机说法正确的是:
(1 分)
多层感知机比感知机只多了一个隐含层
感知机只能形成线性判决平面，无法解决异或问题
多层感知机可以有多个隐含层，但是只能有一个输出单元
隐含层神经元的个数应当小于输入层神经元的个数
单选 3. 多层感知机解决线性不可分问题的原理是:
(1 分)
分而治之，对原始问题空间进行划分
将原始问题向更高维空间映射
在输出层和隐含层之间形成非线性的分界面
将原始问题在隐含层映射成线性可分问题

第四节

1. 在误差逆传播算法中，输出层神经元权重的调整机制和感知机的学习规则相比:
(1 分)
考虑到线性不可分问题，学习规则更为复杂
一模一样，等价于多个感知机
遵循相同的原理，激励函数可能有所不同
所有输出层神经元的权重需要同步调整
单选 2. 在误差逆传播算法中，隐含层节点的误差信息应当:
(1 分)
根据自身的期望输出和实际输出的差值计算
根据所有输出层神经元的误差的均值计算
根据自身下游神经元的误差进行加权计算
根据自身下游神经元的误差的均值计算

单选 3. 为了克服学习空间中存在的局部最优点应当：
（1 分）
尝试从不同的初始点开始训练
将权重初始化为接近于 0 的值
采用较小的学习率
增加隐含层神经元个数
单选 4. 关于学习率参数的设置，正确的描述是：
（1 分）
较大的值有助于提高算法的收敛稳定性
较小的值有助于提高算法的收敛速度
在开始阶段应该较大，然后逐渐减小
在开始阶段应该较小，然后逐渐增大
单选 5. 在权重更新公式中引入冲量的主要目的是：
（1 分）
提高算法的收敛精度
提高算法的稳健性
提高算法的全局优化能力
有助于摆脱误差平缓区域

第五节

1. 前馈神经网络适用的场景为：
（1 分）
训练时间有限
训练样本含有噪声
需要较快的测试响应速度
较好的可解释性
多分类问题
单选 2. 在 Elman 网络中，第 T 时刻网络的输出取决于：
（1 分）
当前的网络输入
当前的网络输入和第 T-1 时刻网络的内部状态
第 T-1 时刻网络的内部状态
当前的网络输入和第 1 到 T-1 时刻网络的内部状态
多选 3. 以下关于 Hopfield 网络特性的描述正确的是：
（1 分）
基于内容的检索
联想记忆功能
误差逆传播
含噪声的模式识别

第六节

1. 视频中的乒乓球运动员是：
(1 分)
瓦尔德内尔
萨姆索诺夫
波尔
刘国梁
单选 2. KUKA 机器人来自以下哪一个国家：
(1 分)
英国
德国
美国
天朝
单选 3. 如果机器人采用神经网络作为控制器，其最有可能的输入是：
(1 分)
击球的声音：麦克风
图像：天花板上的高速摄像头
图像：机械臂上的高速摄像头
球的落点：桌面上的压力传感器

第五章

第一节

1. 在 SVM 领域中，margin 的含义是：
(1 分)
盈利率
马金
间隔
保证金
单选 2. 线性 SVM 和一般线性分类器的区别主要是：
(1 分)
是否进行了空间映射
是否确保间隔最大化
是否能处理线性不可分问题
训练误差通常较低
单选 3. 为什么通常要选择 margin 最大的分类器？
(1 分)
所需的支持向量个数最少
计算复杂度最低

训练误差最低
有望获得较低的测试误差
单选 4. 假设超平面为 w*x+b=0，其 margin 的大小为：
（1 分）
1/|w|
2/|w|
|b|/|w|
2|b|/|W|
单选 5. 支持向量（support vectors）指的是：
（1 分）
对原始数据进行采样得到的样本点
决定分类面可以平移的范围的数据点
位于分类面上的点
能够被正确分类的数据点

第二节

1. 在 SVM 的求解过程中，支持向量与 alpha 的关系是：
（1 分）
alpha=0 的数据点是支持向量
alpha>0 的数据点是支持向量
alpha<0 的数据点是支持向量
两者没有固定关系
单选 2. 在 SVM 当中，主要的运算形式是：
（1 分）
向量内积
矩阵乘法
矩阵转置
矩阵分解
单选 3. 软间隔（soft margin）的主要用途是：
（1 分）
解决线性不可分问题
解决不完全线性可分问题
降低算法时间复杂度
提高算法分类精确

第三节

1. 在 SVM 当中进行空间映射的主要目的是：
（1 分）
降低计算复杂度
提取较为重要的特征
对原始数据进行标准化

提高原始问题的可分性

单选 2. 对于 SVM，在映射后的高维空间直接进行计算的主要问题是：

（1 分）

模型可解释性差

计算复杂度高

容易出现奇异矩阵

容易出现稀疏矩阵

单选 3. 所谓 kernel trick，指的是：

（1 分）

利用在原始空间定义的函数替代高维空间的向量内积操作

利用在高维空间定义的函数替代原始空间的向量内积操作

核函数的导数具有简单的解析解，简化了运算

核函数具有固定的上下界，可以输出（-1, +1）区间中的连续值

单选 4. 通过运用核函数，我们可以：

（1 分）

提高算法的可解释性

生成数量较少的支持向量

生成数量较多的支持向量

避免高维空间运算，降低算法复杂度

第四节

1. SVM 核心技术的发展经历了：

（1 分）

10 年

20 年

30 年

40 年

单选 2. 线性 SVM 思想最初被提出的时候，你在：

（1 分）

上幼儿园

上小学

上中学

不知道在哪儿

单选 3. 一个分类模型的 capacity 指的是：

（1 分）

能够解决几分类问题

能解决多大规模的问题

能将多少个点分开，不论如何分配标签

能达到的精确度

单选 4. 为什么当两个模型的训练误差相同或接近的时候，通常会选择比较简单的一个：

（1 分）

复杂模型的测试误差一定较大

简单模型的测试误差一定较小

在相同置信度条件下，复杂模型的测试误差上界较大
只是一种经验，并没有理论依据
多选 5. Владимир Наумович Вапник（Vladimir Vapnik）为什么是真神：
（1 分）
惊天引用次数
支持向量机开天辟地
统计学习理论一代宗师
目光如炬，深不可测

第六章

第一节

1. 聚类中的簇与分类中的类的关系是：
（1 分）
簇即是类、类即是簇
簇是类的一种具体表现形式
类是簇的一种具体表现形式
不是一码事，但实际中有一定联系
单选 2. 在市场营销中，聚类最有可能帮助经营者：
（1 分）
对客户群进行划分
进行商品推荐
识别优质客户
辅助商品定价
多选 3. 一个好的聚类算法应当具备哪些潜质：
（1 分）
能够处理非球形的数据分布
能够处理噪点和离群点
对样本输入序列不敏感
对海量数据的可扩展性
单选 4. 关于数据预处理对聚类分析的影响的错误说法是：
（1 分）
可能改变数据点之间的位置关系
可能改变簇的个数
有助于提升聚类质量
可能产生不确定影响
单选 5. 在基于聚类的图像分割例子中：
（1 分）
色彩越复杂的图，需要的簇的个数越少
属于同一个物体的像素对应同一个簇

簇的个数越少，分割后图像越接近原始图像
簇的个数越多，分割后图像越接近原始图像

第二节

1. 如何衡量聚类的质量：
（1 分）
簇内数据点散布越小越好
簇中心点之间的距离越大越好
簇的个数越小越好
需要考虑数据点间的连通性

单选 2. 对于 Silhouette 图表述正确的是：
（1 分）
每个点的取值范围为[0, 1]
每个点的取值越接近于 0 越好
可以体现出簇的紧凑性
对于离群点，取值可能超过 1

单选 3. 关于 K-Means 算法的表述正确的是：
（1 分）
对数据分布没有特殊的要求
能较好处理噪点和离群点
对初始中心点较为敏感
计算复杂度较高

单选 4. K-Means 算法中的初始中心点：
（1 分）
可随意设置
必须在每个簇的真实中心点的附近
必须足够分散
直接影响算法的收敛结果

单选 5. 在 Sequential Leader 算法中：
（1 分）
需对数据集进行多次遍历
无法人为控制最终聚类的个数
需要事先生成初始中心点
聚类结果可能受数据访问顺序影响

第三节

1. 以 K-Means 算法为例，期望最大化算法中的：
（1 分）
模型参数指的是每个数据点的簇标号
隐含参数指的是每个数据点的簇标号
模型参数指的是簇的个数（即 K 值）

隐含参数指的是簇中心点坐标

单选 2. 在掷硬币的例子中，期望最大化算法的隐含参数指的是：

(1 分)

每组实验中正面朝上的次数

每组实验中选择的硬币

每枚硬币正面朝上的概率

每枚硬币被选中的次数

单选 3. 基于模型的聚类与基于分割的聚类相比：

(1 分)

有更高的精确度

有更低的计算复杂度

有更好的鲁棒性

对数据分布有更好的描述性

单选 4. 在混合高斯模型中，每一个数据点：

(1 分)

只能被某一个高斯生成

可以被所有高斯等概率生成

可以被任一高斯生成但概率可能不等

可以被任一高斯生成且概率由高斯的权重决定

单选 5. 在混合高斯模型中，每个高斯的权重：

(1 分)

可以为负值

相加必须等于 0

相加必须等于 1

须由用户预先设定

第四节

1. 与 K-Means 相比，基于密度的 DBSCAN 的优点不包括：

(1 分)

能妥善处理噪点和离群点

能处理不规则的数据分布

不需要预先设定簇的个数

较低的计算复杂度

单选 2. 在 DBSCAN 中，对数据点类型的划分中不包括：

(1 分)

中心点

核心点

边缘点

噪点

单选 3. 在 DBSCAN 中，对于噪点：

(1 分)

划分到最近的簇

所有噪点单独形成一个簇
直接无视
不做特别区分
单选 4. 在层次型聚类中:
(1 分)
需要用户预先设定聚类的个数
需要用户预先设定聚类个数的范围
对于 N 个数据点，可生成 1 到 N 个簇
对于 N 个数据点，可生成 1 到 N/2 个簇
单选 5. 在层次型聚类中，两个点集之间的距离计算方法通常不包括:
(1 分)
由点集间距离最近的一对点的距离决定
由点集间距离最远的一对点的距离决定
由点集间随机的一对点的距离决定
由点集间所有点的平均距离决定

第七章

第一节

1. 已知梭罗的《瓦尔登湖》和柏拉图的《理想国》经常被同时购买，那么:
(1 分)
《瓦尔登湖》的读者很有可能会买《理想国》
《理想国》的读者很有可能会买《瓦尔登湖》
两本书的读者都很有可能买另一本书
得具体问题具体分析
单选 2. 某人买电脑的预算为 6000 元，最终从电脑城买了一台 8000 元的电脑，学术上的解释是:
(1 分)
交叉销售
向上销售
捆绑销售
被坑了
单选 3. 某人望着一柜子衣服，感觉自己没有衣服穿，遂上街四个小时，购得手袋一只、高跟鞋若干双、帽子一顶、丝巾一条、YSL 口红一只..... 针对以上行为，学术上的解释是:
(1 分)
交叉销售
向上销售
捆绑销售
女人啊
单选 4. 百货商场第一层进门区域通常会布置为:
(1 分)

女装
男装
美食广场
香水、化妆品

第二节

1. 关联规则 X->Y 的支持度等同于{X, Y}的支持度。
(1 分)
正确
错误
判断 2. 关联规则 X->Y 的置信度等价于条件概率 P(Y|X)的值。
(1 分)
正确
错误
判断 3. 关联规则 X->Y 是一条强规则指的是{X, Y}在数据库中频繁出现。
(1 分)
正确
错误
多选 4. 一条有价值的关联规则必须满足：
(1 分)
支持度高：足够频繁
置信度高：足够有说服力
前件和后件交集为空
前件和后件必须包含多个 item

第三节

1. 一条关联规则的置信度只需大于预设的阈值就是有价值的规则。
(1 分)
正确
错误
判断 2. 只要关联规则 X->Y 的置信度大于 Y 自身的概率就是一条有价值的关联规则。
(1 分)
正确
错误
判断 3. 因为关联规则描述的是事件之间的条件概率，因此可以用于推断因果关系。
(1 分)
正确
错误
单选 4. 冰激凌和犯罪的例子说明：
(1 分)
冰激凌含有某种神秘物质能够诱发犯罪

从事违法行为后，人们喜欢吃冰激凌平静一下心情
还记得那些年我们一起学过的条件独立吗？
我的智商该充值了

第四节

1. 用蛮力搜索所有的频繁项集的最大困难在于：
（1 分）
每条交易记录可能很长
数据库可能包含很多条交易记录
候选项集总数过于庞大
存储器 I/O 读写耗时

单选 2. 评价一个学者水平的正确方式是：
（1 分）
项目经费
他人引用次数
头衔（杰青、长江、院士）
获奖（国家、省部级科技奖励）

单选 3. 关于 Apriori 算法说法不正确的是：
（1 分）
所有频繁项集的子集都是频繁的
所有不频繁项集的超集都是不频繁的
对频繁项集的搜索遵循 bottom-up 的原则
最终的输出结果是长度最长的频繁项集

单选 4. 对 Apriori 算法工作原理最贴切的说法是：
（1 分）
空间剪枝
启发式搜索
折半查找
分支定界

第五节

1. 从三个频繁项集 {1, 2}, {1, 3}, {1, 4} 中能生成以下哪个可能频繁的项集：
（1 分）
{1, 2, 3}
{1, 2, 4}
{2, 3, 4}
以上均不正确

单选 2. 在 Apriori 算法中，假设已获得 Lk，则寻找 K+1 频繁项集时应确保：
（1 分）
所有可能频繁的 K+1 项集都在 Lk+1 中
尽可能多的 K+1 项集都在 Ck+1 中

所有可能频繁的 K+1 项集都在 Ck+1 中
我有点乱
单选 3. 以下哪种推荐最为靠谱：
（1 分）
已购买 iPhone 6，推荐 iPhone 6 Plus
已购买小米 5，推荐手机贴膜
已购买华为 Mate 7，推荐备用电池
已购买 Galaxy Note 7，推荐防爆套装
单选 4. 以下哪种推荐最有内涵：
（1 分）
喜欢看《碟中谍 3》，推荐《碟中谍 4》
喜欢看《红楼梦上册》，推荐《红楼梦下册》
喜欢看《辛德勒的名单》，推荐《钢琴家》
喜欢看《红磨坊》，推荐《冷山》

第六节

1. 以下关于序列和项集说法不正确的是：
（1 分）
序列中包含项集
完全相同的项集不能在同一序列中重复出现
序列强调时间上的先后顺序
不同序列对应不同的客户 ID
单选 2. 给定 A，B，C 三件商品，以下哪条序列不正确：
（1 分）
<{A, B, C}>
<{A} {A} {C}>
<{B} {A, A} {C}>
<{A, B} {A, B, C}>
单选 3. 已知序列<{2, 3} {3, 6, 5} {8}>，以下哪条序列不是其子序列：
（1 分）
<{2} {8}>
<{2} {3, 6}>
<{3, 5} {8}>
<{2} {3} {5}>

第七节

1. 未来商店中的智能水果称很难被引入国内超市，原因主要是：
（1 分）
识别精度低
成本高昂
操作繁琐

你懂的

单选 2. 未来商店可以对每一位客户的轨迹进行精准记录，对下图合理的解释是：

（1 分）

女人更具有探索精神

女人更喜欢货比三家

女人方向感相对较弱

我可以笑而不语吗？

单选 3. 另一种关联： 150 盏亮着的电灯，各有一个拉线开关控制（每拉动一次拉线，对应灯的状态翻转一次），被顺序编号为 1，2，3，4，…，150. 若将编号为 3 的倍数的灯的拉线各拉一下，再将编号为 5 的倍数的拉 线各拉一下，拉完后亮着的灯数为几盏？

（1 分）

70

80

90

100

多选 4. 未来商店中展示了以下哪些智能服务：

（1 分）

个性化超市导航

全自动收银台

基于 RFID 的商品识别

机器人服务员

物联网智能冰箱

第八章

第一节

1. 在线广告引擎需要考虑的因素有：

（1 分）

你在哪里？

你在看什么？

你是谁？

历史点击记录

单选 2. 网页中植入的广告对手机用户的影响和对 PC 用户的影响相比：

（1 分）

对手机用户影响大

对 PC 用户影响大

区别不明显

因人而异

多选 3. 以下哪些行为可能意味着用户喜欢某一首歌曲：

（1 分）

单曲循环

经常播放
下载到本地
推荐给好友

单选 4. “Your trash can be someone's treasure.” 这句话的意思是：
（1 分）
人的品味差别很大
人的需求不同
要做好废物回收利用
点石成金、变废为宝

多选 5. 以下属于基于内容的推荐的例子有：
（1 分）
小明喜欢吃草莓，推荐她去品尝草莓奶昔
听闺蜜说《花千骨》好看，于是去追剧
小陈喜欢看《那些年》，推荐她看《致青春》
听说最近某影片很火，遂决定去一探究竟

第二节

1. 多义词的存在会导致信息检索时：
(1 分)
召回率降低
准确率降低
召回率提高
准确率提高

多选 2. 根据 TF-IDF 的机理，一个查询词和文档的相关性强说明：
(1 分)
在该文档中频繁出现
在该文档中极少出现
在其它文档中频繁出现
在其它文档中极少出现

单选 3. 在向量空间模型中，两篇文档接近等价于：
(1 分)
夹角余弦值接近于 1
夹角余弦值接近于 0.5
夹角余弦值接近于 0
夹角余弦值接近于-1

单选 4. 隐含语义分析的数学原理与以下哪一种技术最接近：
(1 分)
SVM
LDA
PCA
SMOTE

单选 5. 近义词的存在会导致信息检索时：
(1 分)

召回率降低
准确率降低
召回率提高
准确率提高

第三节

1. PageRank 的计算依据是：
(1 分)
网页的访问量
网页内容的质量
网页内容的类型
网页与其它网页的关系
单选 2. 在计算一个网页的 PageRank 值时需要考虑：
(1 分)
所有相似的网页
所有与之链接的网页
所有指向它的网页
所有它指向的网页
单选 3. 网络中所有网页的 PageRank 值的总和：
(1 分)
恒等于 1
恒等于 0
随着时间推移越来越大
随着时间推移越来越小
单选 4. 具有高 PageRank 值的网页的特征是：
(1 分)
万千宠爱于一身
万花丛中过，片叶不沾身
花自飘零水自流，一种相思两处闲愁
举杯邀明月，对影成三人
单选 5. 斯坦福大学较早即出售谷歌股票的原因是：
(1 分)
对谷歌未来不看好
没见过那么多钱
问世间钱为何物
点此查看

第四节

1. 以下哪些问题会显著影响协同过滤算法的有效性：
(1 分)
新用户

新商品
虚假评价
灰山羊
多选 2. 在协同过滤算法中，需要考虑哪些客户的信息：
（1 分）
与目标客户兴趣相投的
与目标客户兴趣相反的
所有与目标客户有打分交集的
打分矩阵中的全体用户
多选 3. 打分矩阵的主要特点有：
（1 分）
行数较大
列数较大
较为稀疏
对称矩阵
单选 4. 在基于模型的协同过滤算法中，因为空缺值普遍存在，所以推荐使用：
（1 分）
神经网络
支持向量机
朴素贝叶斯
决策树
单选 5. 假设用户打分为“喜欢”或“不喜欢”，在讲义中的例子里，空缺值的编码应为：
（1 分）
0.5
0
00
11

第五节

1. 我们从 Netflix Prize 的例子中应当体会到：
（1 分）
推荐问题在实际中极具挑战
手快有，手慢无
知识就是财富
算法为王
单选 2. Netflix 公司没有继续举办这一比赛的原因是：
（1 分）
经营不善，难以为继
产生纠纷，被人告了
原本是个噱头，结果玩大了
竞赛而已，对公司没有实际价值
多选 3. 你知道吗，Netflix Prize 的冠军队的算法并没有在实际中应用：
（1 分）

为了满足实际需求，需要对算法进行大量修改
所获得的准确率的提升对实际利润影响很小
公司产业转型，在线视频用户行为与 DVD 租碟用户行为差异很大
对于在线视频，可以获得更多用户行为信息
对于在线视频，推荐的精准性要求不需要那么高

单选 4. 推荐算法在实际应用中最大的挑战是：
（1 分）
冷启动问题
打分矩阵的稀疏问题
打分矩阵的维数问题
人心难测

单选 5. 为什么推荐算法的应用领域很少涉及女性服装？
（1 分）
数据采集困难
隐私保护问题
对服装进行定量描述困难
你上次看到两个女人穿一样的衣服是什么时候？

单选 6. 如果你和梅尔·吉普森一样能够听懂女人的心思，你会选择做：
（1 分）
恋爱指导师
商场导购员
街道居委会主任
嘿嘿，你懂的

单选 7. 如果甲乙两人通常只在周末固定时间通电话，二人的关系最有可能是：
（1 分）
夫妻
同事
母子
情人
哥们

第九章

第一节

1. 关于集成学习的说法正确的有：
（1 分）
团结力量大
尺有所短寸有所长
赢者通吃
一个好汉三个帮

单选 2. 关于集成学习算法的说法正确的是：

(1 分)
一种并行的算法框架
一种串行的算法框架
一类全新的数据挖掘算法
一类将已有算法进行整合的算法

多选 3. 集成学习成功的关键在于:
(1 分)
选择尽可能强悍的基础分类器
选择多样性的基础分类器
采用尽可能多的基础分类器
选择合适的基础分类器权重

第二节

1. 以下哪些措施有助于提高基础分类的多样性:
(1 分)
采用不同的训练集
采用不同类型的算法
采用强的基础分类器
采用不同的训练参数
采用不同的数据特征

单选 2. 关于 Bootstrap 采样正确的说法是:
(1 分)
有放回的采样
无放回的采样
样本大小必须与原样本相同
应尽可能保证各原始数据都出现

多选 3. Bagging 的主要特点有:
(1 分)
各基础分类器并行生成
各基础分类器权重相同
只需要较少的基础分类器
基于 Bootstrap 采样生成训练集

单选 4. 在随机森林中，由于采用了 Bootstrap 采样，因此理论上有多少原始样本没有被选入训练集?
(1 分)
1/2
1/3
1/4
3/4

单选 5. 如何充分利用现有数据评价随机森林的性能:
(1 分)
10-fold 交叉验证
用 OOB 中的数据作为测试集

用不在 OOB 中的数据作为测试集
用所有数据作为测试集

第三节

1. 在基于 Stacking 的集成模型中：
（1 分）
第一层的基础分类器必须采用同一种分类器
在训练第二层分类器时应采用各基础分类器的输出作为输入
在训练第二层分类器时应采用在基础分类器中占多数的输出值作为输出
第二层分类器的作用是对基础分类器的输出进行集成

多选 2. 对 Boosting 模型的描述正确的是：
（1 分）
采用串行训练模式
基础分类器通常应采用强分类器
通过改变训练集进行有针对性的学习
基础分类器采用少数服从多数原则进行集成

多选 3. 对 AdaBoost 描述正确的是：
（1 分）
可以集成出训练误差任意低的分类器
基础分类器可以任意弱（准确率高于 50%）
通过对样本进行加权达到改变训练集的效果
被当前基础分类器分错的样本的权重将会减小

第四节

1. 在 AdaBoost 算法中，Z 的作用是：
（1 分）
确保在 t+1 代所有样本权重之和为 1
一个用于标准化的变量，可有可无
可以用来描述算法的训练误差上界
较小的 Z 值说明当前分类器的效果较好

单选 2. AdaBoost 中核心参数 alpha 的取值为(e 为模型错误率)：
（1 分）
1/2ln((1-e)/e)
ln((1-e)/e)
1/2ln(e/(1-e))
ln(e/(1-e))

多选 3. AdaBoost 算法的优点有：
（1 分）
容易实现
可解释性强
参数选择简单

不容易过学习
抗噪声能力强

第五节

1. AdaBoost 中基础分类器的权重设置策略存在的问题有：
（1 分）
计算复杂
不能保证是最优解
需要用户进行手工设置
不能根据测试样本进行自适应调整
多选 2. 以下对 RegionBoost 算法描述正确的是：
（1 分）
基础分类器权重根据当前输入样本计算得出
每个基础分类器需要一个额外的可信度模型
每个基础分类器的权重针对不同输入样本有所区别
可信度模型用于估计基础分类器对特定输入的准确度
多选 3. RegionBoost 与 AdaBoost 相比：
（1 分）
训练误差通常降低较慢
训练误差能够趋近于 0
测试误差可能优于 AdaBoost
有较多的参数需要设置

第十章

第一节

1. 进化论的基本要素包括：
（1 分）
适者生存
用进废退
杂交变异
迭代优化
单选 2. 在针对植株高度的豌豆实验中，如果把杂种一代（Ts）和纯矮株(ss）杂交，产生的子代中高、矮豌豆的比例应该接近于：
（1 分）
1:1
1:2
2:1
3:1

多选 3. 属于进化计算范畴的算法有：
（1 分）
Simplex Method
Gradient Descent
Genetic Algorithms
Genetic Programming
Gauss–Newton Algorithm
Expectation Maximization Algorithm
单选 4. 蚂蚁寻找最短路径的能力来源于：
（1 分）
个体的超强感知能力
个体的全局规划能力
蚁后的英明领导
群体合作
多选 5. 鸟人飞行尝试以失败告终，其可能的原因有：
（1 分）
你站得不够高，怎么知道自己不是一只鹰?
失败是成功之母，尝试的次数不够多
人体肌肉骨骼特点决定
简单模仿、生搬硬套

第二节

1. 给定 100 个属性，挑选出不超过 3 个属性，其解空间的大小为：
（1 分）
161700
166650
166750
166900
单选 2. 根据 Bremermann's Limit，质量为一公斤的计算机的理论最高运算速度的量级为：
（1 分）
10 的 30 次方
10 的 40 次方
10 的 50 次方
10 的 75 次方
单选 3. 在 Bin Packing 问题中，如果使用 First Fit 方法，依次将物品放入第一个可以容纳的箱子中，那么最终使用的箱子的个数的已知上界最接近于（OPT 为理论最优解）：
（1 分）
1.2 OPT
2.0 OPT
2.5 OPT
3.0 OPT
多选 4. 优化问题的难度主要来源于：
（1 分）

问题的高维度
问题变量间的强相关性
解空间中的多极值点
巨大的解空间

单选 5. 对于从某地出发，访问 15 个城市的 TSP 问题，其解空间的大小最接近于：
（1 分）
1,000,000,000,000
100,000,000,000
10,000,000,000
1,000,000,000

第三节

1. 进化计算实现全局优化能力的途径有：
（1 分）
基于种群，减小初始点影响
并行搜索，疏而不漏
动态调整，合理分配搜索资源
交叉变异，优势互补

多选 2. 进化计算领域的三个主要会议是：
（1 分）
CEC
GECCO
IJCNN
PPSN

多选 3. Blondie24 中运用到的技术包括：
（1 分）
神经网络
协同进化
比赛对局数据库
专家知识库

多选 4. 以下关于进化计算说法不正确的是：
（1 分）
要求目标函数可导
要求目标函数为凸函数
要求目标函数的定义域为凸集
要求目标函数有明确的数学表达式

第四节

1. 以下哪种个体选择策略容易造成“赢者通吃”现象：
（1 分）
Rank Selection

Rouletee Wheel Selection
Tournament Selection
Truncation Selection

单选 2. 下图中哪一个对应格雷码?
（1 分）
我选左边
我选右边
臣妾真的不知道

多选 3. 在 Tournament Selection 中，每次参与 PK 的个体越多：
（1 分）
强势个体受益越大
强势个体受益越小
弱势个体受益越大
弱势个体受益越小

多选 4. 以下哪些措施有助于保持遗传算法搜索过程的稳定性：
（1 分）
Elitism
(μ+λ) Strategy
采用较大的种群
采用较高的变异率

第五节

1. 关于杂交算子说法正确的是：
（1 分）
有助于保持种群的基因多样性
遗传算法的主要搜索方式
通过基因重组生成新的个体
体现出对现有搜索结果的精细利用（Exploitation）

多选 2. 关于变异算子说法正确的是：
（1 分）
有助于保持种群的基因多样性
通常独立作用于个体的某一位基因
设置较大的变异率有助于提高收敛速度
体现出对解空间的各个区域的探索（Exploration）

多选 3. 关于选择算子说法正确的是：
（1 分）
不影响种群的基因多样性
可视为搜索资源分配的调节机制
进化初期 Selection Pressure 过大易导致不成熟收敛
进化初期 Selection Pressure 过大易导致算法收敛过慢

单选 4. No Free Lunch 定理的寓意是：
（1 分）
不能白吃白占

吃人嘴短、拿人手短
不存在所谓的最优算法
贪小便宜吃大亏
多选 5. 遗传算法的参数主要包括：
(1 分)
杂交率
变异率
种群大小
选择算子
多选 6. 遗传算法参数调整的方法有：
(1 分)
可随机设置
根据经验设置固定值
设计启发式机制，动态调整参数值
针对具体问题，利用少量试验寻找合适的参数值
单选 7. 假设优化函数分别为 F 和 G，则对于 Pareto Front 上的任意两个点 x 和 y：
(1 分)
F(x)=F(y) 且 G(x)=G(y)
若 F(x)小于 F(y)，则 G(x)大于 G(y)
若 F(x)小于 F(y)，则 G(x)小于 G(y)
若 F(x)小于 F(y)，则 G(x)=G(y)

第六节

1. 遗传程序设计中个体的表现形式为：
(1 分)
树
图
表
向量
单选 2. 遗传程序设计中个体的大小：
(1 分)
固定不变
可能随杂交操作变化
可能随变异操作变化
随着迭代逐渐增大
多选 3. 在遗传程序设计中对两个完全相同的个体进行杂交：
(1 分)
没有意义，因为不能产生新的基因
没有意义，因为子代和父代完全相同
有意义，有可能生成新的个体
有意义，因为杂交点的选取可以不对称

第七节

1. 在 Evolutionary Arts 中，最有挑战性的环节是：
（1 分）
遗传算子
适应度函数
编码表示
计算复杂度
多选 2. 关于进化计算的描述正确的是：
（1 分）
可用于模拟和分析自然界的进化过程
可用于解决各类工程优化问题
包含各种从自然中获得灵感的算法
万物皆进化
多选 3. 一个真正智能的系统应具有以下特征：
（1 分）
自学习
可进化
不需要人的具体指导
能够超过设计者的能力范畴

第十一章

第一节

1. 这是一张 60 年前的照片，画面中的女人在展示当时最先进的 IBM 305 RAMAC 商用计算机，配备容量为 3.75 MB 的 IBM 305 硬盘存储单元（左侧圆柱形装置）。该系统当时每月租金按目前价格折算最接近于：
（1 分）
1 万美元
2 万美元
3 万美元
5 万美元
单选 2. 针对在社交媒体频繁发自拍的行为说法正确的是：
（1 分）
这是病，得治
招人烦，容易被拉黑
真可怜，连个拍照的人都没有
可能不经意间泄露个人隐私，造成安全隐患
我嚼着以上都对
多选 3. 在社交平台上点赞的行为可能泄露自己的：

（1 分）

性别

年龄

职业

感情状况

取向

单选 4. 目前，数据的被遗忘权主要指的是：

（1 分）

我有权忘记自己的过去

我有权删除自己保存的数据

我有权要求网站直接删除和自己有关的数据

我有权要求搜索引擎删除过时的和自己有关的搜索结果

单选 5. 数据的可携带权指的是：

（1 分）

公民享有随身携带数据的权力

公民享有访问政府公开数据的权力

公民享有要求服务商提供个人数据迁移便利的权力

公民享有以不同可移动媒介存储个人数据的权力

单选 6. 在美国，目前每年拘禁在监狱和拘留所等机构的人数最接近于：

（1 分）

一百万

两百万

五百万

七百万

单选 7. 利用大数据分析技术进行预防犯罪：

（1 分）

可保天下无贼、国泰民安

主要难点在于选择合适的预测模型

人心难测，难于上青天

可针对群体进行防范，但不宜针对个人

单选 8. 《少数派报告》的主演是：

（1 分）

汤姆·汉克斯

汤姆·克鲁斯

乔治·布鲁尼

皮尔斯·布鲁斯南

莱昂纳多·迪卡普里奥

单选 9. 何老师的手表是以下哪个品牌：

（1 分）

Apple Watch

Breitling

Cartier

Hermès

OMEGA

Rolex